Real Time Tennis Match Prediction Using Machine Learning

CS229 Fall 2017 Project Proposal

Team members:Yang "Eddie" Chen (yc4@), Yubo Tian(yubotian@), Yi Zhong (yizhon@)

1. Introduction

1.1. Overview

Tennis matches are fun to watch because they are full of surprises. In the 2017 Stuttgart Open, Roger Federer, then an 18-time grand slam champion, was beaten by world No. 302 Tommy Haas. Federer lost the opening match at a grass-court tournament, which hadn't happened since 2002. How can we predict a rare loss like this? Based on past performance alone, any model would have predicted a Federer win in pre-game bets.

This motivates us to apply machine learning to predict tennis matches in real time; in particular, we want to explore how well we can predict match results, and then set-by-set outcome (and maybe even more granularly). The hope is that although perhaps our model would still have Federer as the winner before game, once we feed it real time data, it would be able to call the next set's outcome and predict the underdog victory before it happens.

1.2. Background

Tennis matches have a hierarchical order: point -> game -> set -> match[1]. We can predict the probability of winning a point, a game, a set or a match. While we ultimately want to predict the final outcome, predicting the probability of winning a point, a game or a set may be more accurate, since players often start a new set with different strategy.

1.3. Motivation for In-Game Prediction

- Extensive researches have been done on tennis pre-game predictions. ATP provides prediction before game based on past performance (including previous encounter, current ranking, etc)[2]

[1] http://www.tennistips.org/tennis-scoring.html

[2] http://www.atpworldtour.com/en/players/fedex-head-2-head/rafael-nadal-vs-grigor-dimitrov/N409/D875

- Anecdotally, most of us probably have been consumers of ESPN's in-game predictions when watching an NFL game on ESPN's website, with a win chance chart displaying the real-time probability that ESPN thinks each team has.
- As pointed out by [6.1], an in-game "approach to tennis match prediction can be more accurate, as it allows the model to capture the change in a player's performance over the course of the match. For example, different players fatigue during a match in different ways. " It's also surprising that not a lot have been done for **real-time, in-game predictions** - we stand a chance to build upon some solid work in an innovative direction
- Results from prediction can be extended to give real-time coaching advice to support game strategy decision.
- Selected features from pre-game prediction can be used in in-game prediction as well, which accounts for both current play and past play and gives us a good starting point.

2. Model Representation and Intended Experiments

We plan to start by exploring the methods outlined in [6.1] and predict the final outcome of a match based on past plays. Then, we want to predict more granularly the outcome of matches, sets, games and points based on past plays and real-time performance data, in that order.

Output (Y) is **win or loss**, reducing this to a classification problem. We have Y from the past since we know the exact outcome of a historical match. We can evaluate the accuracy of different methods by comparing our prediction against the actual results (following a training-validation-test split). Input (X) includes features for prediction.

For all methods we intend to use, we will need to do feature selection to fine tune models.

3. Methods for Classification

We plan to use different methods for classification and evaluate the error rate, pros and cons. Potential methods include:

- Logistic Regression
- Decision Tree
- Random Forest
- GLM
- Support Vector Machine (SVM)
- Ensemble (like bagging and boosting) and other techniques to fine tune the abovementioned models

4. **Feature Construction** [source data available in [5]]

4.1. Pre-Match Stats: we will heavily leverage findings from [6.1] to select features for pre-match stats; see the Appendix for prominent features from [6.1].

4.2. On Court Stats: we plan to leverage findings from pre-match stats, and find deviation in performance. This may include computing a metric for:

- Total points won
- Total winners
- Time spent on each game
- Number of unforced errors
- Serve statistics
- Fatigue
- Other game performance metrics
- Etc.

5. Data

Jeff Sackmann has been driving a crowdsourcing project to document point by point data for pro tennis matches, available on GitHub[3]. Details can be found in the Appendix.

6. Related Prior Work

6.1. Machine Learning for the Prediction of Professional Tennis Matches

This paper uses machine learning to predict final outcome. The author spent a lot of effort on feature selection to make prediction based on pre-game data. We intend to leverage his finding and extend to **in-game prediction**.

https://www.doc.ic.ac.uk/teaching/distinguished-projects/2015/m.sipko.pdf

6.2. Real-time prediction to support decision-making in soccer

http://ieeexplore.ieee.org/document/7526923/?reload=true

7. Potential Extension

Future extension of this project include many aspects. From the prediction model, we can see what features have significant effect on the probability of winning and give advice accordingly. In addition, this can be generalized for different sports. Tennis has one of the most complicated scoring structure, whereas games like basketball or soccer simply cumulate scores. Most of the prominent features may apply across sports, and hence make our prediction model extensible.

[3] https://github.com/JeffSackmann

Appendix

Data Source and Scheme Details

- *Tennis_pointbypoint*[4]
  - Sequential point-by-point data for tens of thousands of pro matches
  - Each row in these files represents one match, and contains the following:
    - Date
    - tour (ATP/CH[allenger]/FU[tures]/WTA/ITF [women])
    - draw (Main [draw] or Qual[ifying])
    - server1 (the player who served first)
    - server2
    - winner (1 or 2, corresponding to one of the previous two columns)
    - pbp (S(erver won), R(returner won), A(ce), D(ouble fault))
    - score
    - adf_flag (1 if the point sequence notes any aces or double faults, 0 if not; see below)
- *Tennis_atp*[5] */ Tennis_wta*[6]
  - ATP(men)/WTA(women) player file, plus historical rankings
  - Description:
    - Player files: player_id, first_name, last_name, hand, birth_date, country_code.
    - Ranking files: ranking_date, ranking, player_id, ranking_points.

- *Tennis Match Charting Project*[7]
  - Detailed shot-by-shot records of 3k professional matches
  - Description:
    - MCP match records contain shot-by-shot data for every point of a match, including the type of shot, direction of shot, depth of returns, types of errors, and more.
    - player/ match-aggregated data also available at:
      - http://www.tennisabstract.com/charting/meta.html
- *Tennis Explorer: Injured players*[8]

[4] https://github.com/JeffSackmann/tennis_pointbypoint
[5] https://github.com/JeffSackmann/tennis_atp
[6] https://github.com/JeffSackmann/tennis_wta
[7] https://github.com/JeffSackmann/tennis_MatchChartingProject
[8] http://www.tennisexplorer.com/list-players/injured/

- Up-to-date injury report for tennis players
- Description:
    - Star_date, player_name, tournament, injury reason

Prominent features listed in [6.1]

- Player details:
    - Ranking
    - Injury/ Retirement
- Head-to-Head Details
    - Age
    - Hand
    - Head-to-head win rate
    - Win rate against common opponents
- Match Details
    - Surface
    - Prize Money
    - Odds Predicted by XXX (other predictions)